A sequential Monte Carlo approach for MLE in a plant growth model

Samis Trevezas, Paul-Henry Cournède

A sequential Monte Carlo approach for MLE in a plant growth model

Samis Trevezas[1,2‡] and Paul-Henry Cournède[1,2§]

accepted in Journal of Agricultural, Biological, and Environmental Statistics, 27 february 2013

Ecole Centrale Paris, Laboratoire MAS, Châtenay Malabry, F-92295, France
INRIA Saclay, Île-de-France, EPI Digiplante, Orsay, F91893, France

Abstract

Parametric identification of plant growth models formalized as discrete dynamical systems is a challenging problem due to specific data acquisition (system observation is generally done with destructive measurements), non-linear dynamics, model uncertainties and high-dimensional parameter space. In this study, we present a novel idea of modeling plant growth in the framework of non-homogeneous hidden Markov models ([6]), for a certain class of plants with known organogenesis (structural development). Unknown parameters of the models are estimated via a stochastic variant of a generalised EM (Expectation-Maximization) algorithm and approximate confidence intervals are given via parametric bootstrap. The complexity of the model makes both the E-step and the M-step non-explicit. For this reason, the E-step is approximated via a sequential Monte-Carlo procedure (sequential importance sampling with resampling) and the M-step is separated into two steps (Conditional-Maximization), where before applying a numerical maximization procedure (quasi-Newton type), a large subset of unknown parameters is updated explicitly conditioned on the other subset. A simulation study and a case-study with real data from the sugar-beet are considered and a model comparison is performed based on these data. Appendices are available online.

keywords: dynamical system; ECM algorithm; maximum likelihood estimation; parametric identification; plant growth model; sequential Monte-Carlo

1 Introduction

The current study is motivated by the need of a deeper understanding of individual plant growth dynamics. On one hand, such knowledge could serve as the basis for simplified but satisfying descriptions of the interactions among complex ecophysiological phenomena which guide plant growth. On the other hand, it is a standpoint for improving population based models in the direction of a better prediction of yields in fields or greenhouses at a global scale. A general family of models of Carbon allocation formalized as dynamic systems serves as the basis for our study. They belong to the class of functional-structural plant models (FSPMs, [29]), which combine the description of both structural development and ecophysiological functioning. A generic model of this kind is the GreenLab model introduced by [11]. In [10], a first approach for parameter estimation was introduced but based on the rather restrictive assumption of

‡Corresponding author; e–mail: samis.trevezas@ecp.fr
§e–mail: paul-henry.cournede@ecp.fr

an underlying deterministic model of biomass production and uncorrelated errors in the mass measurements of different types of organs in the plant structure.

The objective of this paper is to introduce a common framework for statistical analysis in a large variety of plant species by taking into account process and measurement errors. We provide a frequentist-based statistical methodology for state and parameter estimation in plants with deterministic organogenesis (structural development) rules. A lot of agronomic plants can be modeled this way, from maize [16] to rapeseed [20], but trees as well [25]. This framework can serve as the basis for statistical analysis in plants with more complex organogenesis, see e.g., [24].

The rest of this article is organized as follows. In the next section, we describe the version of the GreenLab model of plant growth which corresponds to the deterministic organogenesis assumption. In Section 3, by taking into account process and measurement errors, we show how plant growth models of this type can be represented as non-homogeneous hidden Markov models. In Section 4, we present an appropriate state estimation technique based on sequential importance sampling with resampling and how it can be used for performing maximum likelihood estimation via a stochastic variant of the Expectation Conditional-Maximization algorithm. In Section 5, a simulated case evaluates the performance of our algorithm and a case-study based on real measurements of sugar beet plant tests its ability in a real-world scenario. A model comparison is also performed. This article ends with a discussion on the results of this study and some perspectives for further work.

2 Preliminaries and Description of the GreenLab model

A large variety of crop models or plant growth models are formulated as source-sink models ([31]) to describe Carbon allocation between different compartments of plant structure: sinks (all organs) compete for the biomass produced by sources (generally leaves). The specificity of the GreenLab model introduced by [11] lies in its mathematical formulation as a nonlinear discrete dynamic system, see [9]. The discretization takes place by partitioning growth into growth cycles (GCs) with duration depending on a certain amount of accumulated heat which triggers the appearance of elementary botanical entities, called phytomers ([1]). The set of different types (here classes) of plant organs (e.g., blade, petiole, internode, flower, ...) is denoted by $\mathcal{O}$ and depends on the plant under study. For example, in the sugar-beet $\mathcal{O} = \{\text{blade}, \text{petiole}, \text{root}\}$, or symbolically $\mathcal{O} = \{b, p, r\}$. Each GC is determined by the following mechanisms: organogenesis and functioning. Organogenesis is the procedure of creation of new organs in the plant. In this study, we restrict ourselves to the case where organogenesis is known and deterministic. So, the interest will be focused on functioning represented as a recursive mechanistic procedure of biomass allocation and production.

2.1 Biomass allocation

Starting by the initial mass q_0 of the seed, at the n-th GC a certain amount q_n of biomass is produced and is available for distribution to all expanding organs. In order to determine the allocation pattern we need to make some assumptions. Let t_o be the expansion time period and p_{al}^o an unknown euclidean vector of fixed parameters for each class of organs $o \in \mathcal{O}$. All allocation parameters are contained in the vector $p_{al} \stackrel{\text{def}}{=} (p_{al}^o)_{o\in\mathcal{O}}$.

Assumption 1. *i) At $GC(n)$ the produced biomass q_n is fully available for allocation to all*

expanding (preexisting + newly created) organs and it is distributed proportionally to the class-dependent empirical sink functions given by

$$s_o(i\,;p_{al}^o) = p_o\,c\left(\frac{i+0.5}{t_o}\right)^{a_o-1}\left(1-\frac{i+0.5}{t_o}\right)^{b_o-1}, \quad i\in\{0,1,\ldots,t_o-1\}, \tag{1}$$

where $p_{al}^o = (p_o, a_o, b_o) \in \mathbb{R}_+^* \times [1,+\infty)^2$ *and* $(p_o)_{o\in\mathcal{O}}$ *is a vector of proportionality constants representing the sink strength of each class (by convention* $p_b = 1$*) and* c *is the normalizing constant of a discrete* $Beta(a_o, b_o)$ *function, where its unnormalized generic term is given by the product of the two last factors of (1).*
ii) $t_o = T$ *for all* $o \in \mathcal{O} - \{r\}$*, where* T *denotes their common value.*

Assumption 1(ii) is only used to simplify the subsequent mathematical notations and it can be relaxed. Concerning the parameterization, some of the aforementioned parameters could be considered fixed for identifiability reasons (depending on the number of available data), therefore reducing the dimension of the parameter vector.

In order to determine the percentage of biomass which is allocated to each expanding organ at each GC we need to make explicit the associated normalization constant.

Definition 1. *The total biomass demand at* $GC(n)$*, denoted by* d_n*, is the quantity expressing the sum of sink values of all expanding organs at* $GC(n)$*.*

Let us denote by $\{(N_n^o)_{o\in\mathcal{O}}\}_{n\in\mathbb{N}}$ the vector-valued sequence of preformed organs at each GC (plant organogenesis specific and deterministic in this study). It is straightforward to see by Assumption 1, by the fact that an organ is in its i-th expansion stage if and only if (iff) it has been preformed at $GC(n-i)$, and by the uniqueness of the root that

$$d_n(p_{al}) \;=\; \sum_{o\in\mathcal{O}-\{r\}}\;\sum_{i=0}^{\min(n,T-1)} N_{n-i}^o\, s_o(i;p_{al}^o) + s_r(n;p_{al}^r). \tag{2}$$

2.2 Biomass production

Except for the initial mass of the seed q_0 subsequent biomasses $\{q_n\}_{n\geq 1}$ are the result of photosynthesis and leaf blades are the only organs to participate in this procedure. At a given GC the total photosynthetically active leaf blade surface is formed by summing the surface of leaf blades which are photosynthetically active at the same GC. In this direction, we have the following definition.

Definition 2. *i) The photosynthetically active blade surface at* $GC(n+1)$*, denoted by* s_n^{act}*, is the quantity expressing the total surface area of all leaf blades that have been preformed until* $GC(n)$ *and will be photosynthetically active at* $GC(n+1)$*,*
ii) the ratio (percentage) of the allocated biomass Q_l *which contributes to* s_n^{act} *will be denoted by* $\pi_{l,n}^{act}$*.*

For the rest of this article we make the following assumption.

Assumption 2. *i) The initial mass of the seed* q_0 *is assumed to be fixed and known,*
ii) the leaf blades have a common photosynthetically active period and equals T*,*
iii) the leaf blades have a common surfacic mass denoted by e_b*.*

By Assumption 2(ii) and Definitions 1 and 2(i), we have that a blade being in its j-th expansion stage at $GC(l)$ contributes to s_n^{act}, iff $l + T - j \geq n + 1$. By combining this with Definition 2(ii), we get the following parametric expression:

$$\pi_{l,n}^{act}(p_{al}) = \frac{1}{d_l(p_{al})} \sum_{j=0}^{\min(l,l+T-n-1)} N_{l-j}^b s_b(j; p_{al}^b), \qquad (n-T+1)^+ \leq l \leq n, \tag{3}$$

where $x^+ = \max(0, x)$, d_l is given by (2), s_b by (1) and N_n^b is the number of preformed blades at $GC(n)$. By Assumption 2(ii)-(iii) and Definition 2, an expression of s_n^{act} is obtained by dividing the mass of the active blade surface by its surfacic mass, to get:

$$s_n^{act}(q_{(n-T+1)^+:n}; p_{al}) = e_b^{-1} \sum_{l=(n-T+1)^+}^{n} \pi_{l,n}^{act}(p_{al})\, q_l, \tag{4}$$

where $\pi_{l,n}^b(p_{al})$ is given by (3) and $x_{i:j} = (x_i, \ldots, x_j)$, for $i \leq j$ and x a generic variable. Now, we describe how $\{q_n\}_{n\geq 1}$ is obtained.

Assumption 3. *In the absence of modeling errors, the sequence of produced biomasses $\{q_n\}_{n\geq 1}$ is determined by the following recurrence relation known as the empirical Beer-Lambert law (see [16]):*

$$q_{n+1} = F_n(q_{(n-T+1)^+:n}, u_n; p) = u_n\, \mu\, s^{pr} \left\{ 1 - \exp\left(-k_B \frac{s_n^{act}(q_{(n-T+1)^+:n}; p_{al})}{s^{pr}} \right) \right\}, \tag{5}$$

where u_n denotes the product of the photosynthetically active radiation during $GC(n)$ modulated by a function of the soil water content, μ is the radiation use efficiency, s^{pr} is a characteristic surface that represents the two-dimensional projection on the ground, of space potentially occupied by the plant, k_B is the extinction coefficient in the Beer-Lambert extinction law, s_n^{act} is given by (4) and $p \overset{def}{=} (\mu, s^{pr}, k_B, p_{al})$.

Note that q_{n+1} also depends on p_{al}, but only through s_n^{act}, and that p could have lower dimension if some of the aforementioned parameters are fixed or calibrated in the field.

2.3 Parameter estimation

In [10] a parameter identification method for individual plant growth was proposed based on the GreenLab formulation described as above. The available data Y contain organ masses, measured at a given $GC(N)$ by censoring plant's evolution (destructive measurements). The authors define a multidimensional state sequence which records the theoretical masses of all organs present in the plant at $GC(N)$ and make the following assumptions: i) no errors exist in the production equation given by (5), ii) measurements errors are independent normal distributions, sharing common variance parameters ϕ_o for common classes of organs. If we denote by X_N the vector of all produced biomasses until $GC(N)$ and by $\phi = (\phi_o)_{o\in\mathcal{O}}$, then by these assumptions we have

$$Y \sim \mathcal{N}_d(G(X_N; p), \Sigma(\phi)), \tag{6}$$

where G is the d-dimensional allocation function (parameterized by p) which determines the vector of theoretical masses of the-d-observed organs at $GC(N)$ as a function of X_N and $\Sigma(\phi)$ is a diagonal covariance matrix (by the independence assumption). The normality assumption

simplifies considerably parameter estimation, but the complexity of the model leads to a difficult inverse problem. An efficient numerical approximation technique for estimating $\theta \overset{\text{def}}{=} (p, \phi)$ is necessary. Maximum likelihood estimation via a Gauss-Newton procedure (see [3]) or an adaptation of the two-stage Aitken estimator (see [30]) were implemented for this purpose, leading to similar results (see [8]). Nevertheless, independence and normality are crude for this application context. A more challenging approach consists in introducing a process error in equation (5), automatically departing from these restrictive assumptions. In the rest of this article, we develop an appropriate statistical methodology for making state and parameter estimation feasible in a more flexible model. In this way, for a given data set it is possible to choose the most appropriate model with the help of model selection techniques.

3 The GreenLab model as a hidden Markov model

In this section we formalize a more flexible GreenLab version by revisiting assumptions i)-ii) given in subsection 2.3. The key idea consists in rearranging the available data Y into subvectors Y_n by taking into account the preformation time of all available organs. This gives us the possibility to treat data sequentially. Each subvector Y_n will contain the masses of the organs which are preformed at $GC(n)$ and consequently appear for the first time at $GC(n+1)$. If we denote by G_n the vector-valued function that expresses the theoretical masses of all the different classes of organs which started their development at $GC(n)$, then by summing the allocated biomass at each expansion stage and Assumption 1 we obtain directly

$$G_n(q_{n:(n+T-1)}; p_{al}) = \left(\sum_{j=0}^{T-1} \frac{q_{j+n}}{d_{j+n}(p_{al})} s_o(j; p_{al}^o) \right)_{o \in O - \{r\}}. \tag{7}$$

The following assumptions determine the stochastic nature of the model.

Assumption 4. *Let $(W_n)_{n\in\mathbb{N}}$ and $(V_n)_{n\in\mathbb{N}}$ two mutually independent sequences of i.i.d. random variables and vectors respectively, independent of Q_0, where $W_n \sim \mathcal{N}(0, \sigma^2)$ and $V_n \sim \mathcal{N}_d(0, \Sigma)$, with Σ an unknown covariance matrix and d the cardinality of $\mathcal{O} - \{r\}$. By setting $N_n^o = 1$, $\forall o \in \mathcal{O} - \{r\}$,*

i) a multiplicative model error determines the hidden state variables:

$$Q_{n+1} = F_n(Q_{(n-T+1)^+:n}; p)(1 + W_n),$$

where F_n is given by (5),

ii) an additive measurement error determines the observed vectors:

$$Y_n = G_n(Q_{n:(n+T-1)}; p_{al}) + V_n, \quad n \geq 0,$$

where G_n is given by (7).

Remark 1. *i) The error in the state process is assumed to be multiplicative and not additive since in our application context biomasses change orders of magnitude,*
ii) the states Q_n represent masses (idem for Y_n) and rigorously take values in $\mathbb{R}_+$, but as is the case in many applications, we consider normal errors, and not distributions with constrained support, in order to simplify the computations.

The functional representation of the above model is that of a state-space model with state sequence $\mathbf{Q}$, and state and observation equation given by Assumption 4. Since the use of conditional distributions is more appropriate for the statistical inference we prefer its equivalent formulation as a hidden Markov model (HMM). For a wide coverage of HMM theory see [6]. Now, we give the type of HMM induced by our assumptions. The proof is direct and will be omitted.

Proposition 1. *Under Assumptions 1-4, the bivariate stochastic process* $(\mathbf{Q}, \mathbf{Y})$ *defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$*, where* $\theta = (p, \sigma^2, \Sigma)$ *can be represented as an HMM, where*

i) the hidden sequence $\mathbf{Q}$*, with values in* $\mathbb{R}_+$*, evolves as a time-inhomogeneous* T*-th order Markov chain with initial distribution* $\mathbb{P}_\theta(Q_0 \in \cdot) = \delta_{q_0}(\cdot)$ *(dirac at* q_0*), where* $q_0 \in \mathbb{R}_+^*$*, and transition dynamics given by*

$$\mathbb{P}_\theta(Q_{n+1} \in \cdot \mid Q_{(n-T+1)^+:n}) \overset{law}{\approx} \mathcal{N}\left(F_n(Q_{(n-T+1)^+:n}; p), \sigma^2 F_n^2(Q_{(n-T+1)^+:n}; p)\right), \quad (8)$$

ii) the observable sequence $\mathbf{Y}$*, with values in* $(\mathbb{R}_+)^d$*, conditioned on* $\mathbf{Q}$ *forms a sequence of conditionally independent random vectors and each* Y_n *given* $\mathbf{Q}$ *depends only on the vector* $Q_{n:(n+T-1)}$ *with conditional distribution given by*

$$\mathbb{P}_\theta(Y_n \in \cdot \mid Q_{n:(n+T-1)}) \overset{law}{\approx} \mathcal{N}_d\left(G_n(Q_{n:(n+T-1)}; p_{al}), \Sigma\right), \quad n \geq 0. \quad (9)$$

Remark 2. *i) We pinpoint the fact that normality in (8) and (9) is only valid approximately since we deal with positive r.v. . In practice, some constraints should be taken into account for the involved variances.*
ii) Obviously, if we define $\tilde{\mathbf{Q}} \overset{def}{=} (Q_{n:n+T-1})_{n\geq 0}$*, then the bidimensional stochastic process* $(\tilde{\mathbf{Q}}, \mathbf{Y})$ *is a time-inhomogeneous first order HMM, but this theoretical simplification has no other practical benefits.*

4 State and Parameter estimation

The first issue that we tackle in this section is state inference, and in particular, estimation of the conditional distribution of the hidden state sequence $Q_{0:N}$ given the observation vector $Y_{0:N}$. This is a smoothing problem, but it is well known in the classical hidden Markov models ([6]) that smoothing can result from a filtering procedure which approximates recursively the conditional distribution of Q_n given the vector observation $Y_{0:n}$ for $n \in \{0, 1, \ldots, N\}$. State estimation could have an interest in its own right, when model parameters are assumed to be known, or it could be a part of the parameter estimation process. This is exactly the case in the maximum likelihood estimation technique that we propose for this model, where the conditional distribution of the hidden states given the observed ones interferes at each E-step (computed under a fixed parameter value) of an appropriate stochastic variant of a generalized EM-algorithm (Expectation-Maximization). The quality of state estimation determines the quality of approximation of the $\mathcal{Q}$-function and consequently the quality of parameter estimation.

4.1 State estimation via sequential importance sampling with resampling

The exact evaluation of $p_\theta(q_{0:N}|y_{0:N})$ is only feasible in finite state space HMMs ([2]) or HMMs that can be represented as linear Gaussian state space models (see [6], chapter 5). This is

not the case in this model where the state-space is continuous and the unobserved process is nonlinear. In order to approximate this conditional distribution we will use sequential importance sampling ([17]) where the target distribution can be approximated recursively with the help of a sequence of intermediary and lower-dimensional sub-target distributions. A final weighted M-sample $\{q_{0:N}^{(i)}, \tilde{w}_N^{(i)}\}_{i=1}^M$ is said to target the $p_\theta(\cdot|y_{0:N})$ in the sense that realizations from the distribution $\sum_{i=1}^M \tilde{w}_N^{(i)} \delta_{q_{0:N}^{(i)}}(\cdot)$ can be (approximately) considered as realizations from the target one, where $\tilde{w}_N^{(i)}$ denotes the normalized importance weight associated to the i-th path $q_{0:N}^{(i)} \stackrel{\text{def}}{=} (q_0^{(i)}, q_1^{(i)}, \ldots, q_N^{(i)})$ at the final $GC(N)$. Since resampling is needed to avoid weight degeneracy ([15]), the above method is referred to as Sequential Importance Sampling with Resampling (SISR). The term particle filter is also extensively used, especially in the engineering and filtering community. For an excellent coverage of the implementation of SISR techniques in HMMs see [6].

The joint smoothing density $p_\theta(q_{0:N}|y_{0:N})$ corresponding to the HMM given by Proposition 1 can be written as:

$$p_\theta(q_{0:N}|y_{0:N}) = p_\theta(q_{0:N}|y_{0:N-T+1}) \frac{p_\theta(y_{N-T+2:N}|q_{N-T+2:N})}{p_\theta(y_{N-T+2:N}|y_{0:N-T+1})},$$

where the first factor can be computed recursively by

$$p_\theta(q_{0:n}|y_{0:n-T+1}) = p_\theta(q_{0:n-1}|y_{0:n-T}) \frac{p_\theta(q_n, y_{n-T+1}|q_{n-T:n-1})}{p_\theta(y_{n-T+1}|y_{0:n-T})}, \quad T \leq n \leq N.$$

Note that the denominator can be considered as a normalization constant. Indeed, it corresponds to a ratio of successive likelihood terms which are common to all particle paths. The self-normalized versions of IS estimates are invariant to normalization constants and consequently the denominators can be excluded from the IS procedure. This is a great benefit in this context since they cannot be computed explicitly.

Now, we give a description of the SISR algorithm corresponding to the HMM given by Proposition 1. It consists in a modification of the SISR algorithm for first-order (homogeneous) HMMs given for example in Chapter 7 of [6] (Algorithm 7.3.4). This modification takes into account the T-order dependency in the underlying Markov chain. The underlying assumption of time-inhomogeneity both in the hidden Markov chain and in the observation densities adds a further dependency on time but does not change the computational methodology. Since we are dealing with measures that have densities with respect to the Lebesgue measure (or trivially Dirac at the index 0), we use densities and not Radon-Nikodym derivatives in the presentation of our results. Likewise, we will use the term instrumental (or importance) transition density for the conditional density function of Q_n corresponding to the instrumental (or importance) kernel given $Q_{n-T:n-1} = q_{n-T:n-1}$. It will be denoted by $r_\theta(\cdot|q_{n-T:n-1})$ and it could depend on n since we are in a non-homogeneous context. The conditional densities $p_\theta(q_n|q_{n-T:n-1})$ and $p_\theta(y_n|q_{n:n+T-1})$ correspond to the densities of the normal distributions given in Proposition 1 by (8) and (9) respectively.

Algorithm 1. *(SISR corresponding to the HMM given by Proposition 1)*

***Initialization**:*

– *Draw* $\{\tilde{q}_{0:T-1}^{(i)}\}_{i=1}^M$, *where* $\tilde{q}_0^{(i)} = q_0$, $\tilde{q}_n^{(i)} \sim r_\theta(\cdot|\tilde{q}_{0:n-1}^{(i)})$, $n = 1, \ldots, T-1$.

– *Compute the first importance weights*

$$w_0^{(i)} = p_\theta(y_0|\tilde{q}_{0:T-1}^{(i)}) \prod_{n=1}^{T-1} \frac{p_\theta(\tilde{q}_n^{(i)}|\tilde{q}_{0:n-1}^{(i)})}{r_\theta(\tilde{q}_n^{(i)}|q_{0:n-1}^{(i)})}, \quad i = 1, \ldots, M. \tag{10}$$

– *Do resampling if necessary and update trajectories (see General step).*

***General step**: For $n = T, \ldots, N$,*

– *Draw $\{\tilde{q}_n^{(i)}\}_{i=1}^M$, where $\tilde{q}_n^{(i)} \sim r_\theta(\cdot|q_{n-T:n-1}^{(i)})$.*

– *Update the importance weights: For $i = 1, \ldots, M$,*

$$w_{n-T+1}^{(i)} = \begin{cases} w_{n-T}^{(i)}\, p_\theta(y_{n-T+1}|q_{n-T+1:n-1}^{(i)}, \tilde{q}_n^{(i)})\, \dfrac{p_\theta(\tilde{q}_n^{(i)}|q_{n-T:n-1}^{(i)})}{r_\theta(\tilde{q}_n^{(i)}|q_{n-T:n-1}^{(i)})} \\ \qquad\qquad \text{if } T \le n < N, \quad \text{and if } n{=}N, \\ w_{N-T}^{(i)} \displaystyle\prod_{n=N-T+1}^{N} p_\theta(y_n|q_{n:N-1}^{(i)}, \tilde{q}_N^{(i)}) \dfrac{p_\theta(\tilde{q}_N^{(i)}|q_{N-T:N-1}^{(i)})}{r_\theta(\tilde{q}_N^{(i)}|q_{N-T:N-1}^{(i)})} \end{cases} \tag{11}$$

– *Do resampling if necessary (we describe the multinomial resampling): Draw a multinomially distributed random vector with probabilities of success given by the normalized importance weights, i.e., $(N_1, \ldots, N_M) \sim \mathcal{M}(M, \tilde{w}_{n-T+1}^{(1)}, \ldots, \tilde{w}_{n-T+1}^{(M)})$, where $\sum_{j=1}^M N_j = M$ and set for $i = 1, \ldots, M$,*

$$I_n^{(i)} = l, \quad \text{where } l \ge 1 \text{ is such that } \sum_{j=0}^{l-1} N_j < i \le \sum_{j=0}^{l} N_j. \tag{12}$$

Then, set $w_{n-T+1}^{(i)} = c$, where c is a positive constant.

– *Update the trajectories: For $i = 1, \ldots, M$,*

$$q_{0:n}^{(i)} = \left(q_{0:n-1}^{(I_n^{(i)})}, \tilde{q}_n^{(I_n^{(i)})} \right),$$

where $I_n^{(i)} = i$, if there is no resampling, otherwise, it is given by (12).

We obtain two important special cases of Algorithm 1 by specializing the importance transition densities $r_\theta(\cdot|q_{n-T:n-1})$ of Q_n given the history of the hidden process.

- Bootstrap (or Blind) filter: $r_\theta(\cdot|q_{(n-T)^+:n-1}) = p_\theta(\cdot|q_{(n-T)^+:n-1})$. By putting the ratio $p_\theta/r_\theta = 1$ in (10) and (11), we get the first importance weights and the update equation respectively.

- Improved filter:

$$r_\theta(\cdot|q_{(n-T)^+:n-1}) = \begin{cases} p_\theta(\cdot|q_{0:n-1}) & \text{if } 1 \le n \le T-1, \\ p_\theta(\cdot|q_{n-T:n-1}, y_{n-T+1}) & \text{if } T \le n \le N, \end{cases} \tag{13}$$

where $p_\theta(\cdot|q_{n-T:n-1}, y_{n-T+1})$ is a density of a normal distribution, with parameters given in Appendix A. By (11) and (13) we get the following recurrence for the importance weights, for $T \leq n < N$ and $n = N$ respectively:

$$w_{n-T+1}^{(i)} = \begin{cases} w_{n-T}^{(i)} \, p_\theta(y_{n-T+1}|q_{n-T:n-1}^{(i)}) \\ w_{N-T}^{(i)} \, p_\theta(y_{N-T+1}|q_{N-T:N-1}^{(i)}) \prod_{n=N-T+2}^{N} p_\theta(y_n|q_{n:N-1}^{(i)}, \tilde{q}_N^{(i)}), \end{cases} \tag{14}$$

where the initial weights $w_0^{(i)}$ are given as in the bootstrap filter and $p_\theta(y_{n-T+1}|q_{n-T:n-1})$ is a computable incremental weight (see Appendix A).

The term improved filter is used to indicate that choosing r_θ in this way allows to take into account during simulation information from the available data, while the bootstrap filter does not. Note that this filter does not coincide with the optimal filter of [34] since the T-dependence makes the model depart from the assumptions of the latter filter.

4.2 Maximum Likelihood Estimation

The maximum likelihood estimator (MLE) cannot be derived explicitly in hidden Markov models and for this reason an EM-type (Expectation-Maximization) algorithm is a sensible choice (see [13] for the general formulation and [2] as a method for estimation in finite state space HMMs), since it is particularly adapted to incomplete data problems. Starting from an arbitrary initial value, the original deterministic version of the EM-algorithm produces a convergent sequence of parameter updates, under some regularity conditions ([4], [33]). Each iteration of the EM algorithm consists in two steps, the expectation step (E-step) and the maximization step (M-step). In the first step, the conditional expectation of the complete data log-likelihood given the observed data is computed under the current parameter estimate (called Q-function). In the second step, the parameters are updated by maximizing the Q-function of the E-step. For an extensive literature on the EM, see [26] and the references therein.

If the integral involved in the E–step is analytically intractable, then one should approximate the Q-function. This is one of the motivations of a large class of stochastic EM-type algorithms, including in increasing order of generalization, the Stochastic EM (SEM) ([7]), the Monte Carlo EM (MCEM) ([32]) and the Stochastic Approximation EM (SAEM) ([12]). Their common characteristic is that at each iteration they all approximate the Q–function by simulating the hidden state sequence from its conditional distribution given the observed data (see also [19]). In the general context of the EM-algorithm [19] distinguishes three different simulation paradigms: i.i.d. simulation via rejection sampling, independent simulation via importance sampling (IS), or generating dependent samples via Markov chain Monte-Carlo (MCMC). In the context of hidden Markov models the latter two are the most appropriate and [6] give a systematic development of the SISREM (the most appropriate variant of IS in hidden Markov models) and MCMCEM (with explicit M-step), illustrating their performance in the stochastic volatility model of [18] (see Example 11.1.2). Both simulation methods were shown to be similar in their results. For the use of sequential Monte Carlo methods for smoothing in nonlinear state space models with applications to MLE see [28]. The authors use a modified version of a SISREM algorithm based on a fixed-lag technique presented by [21] in order to robustify the parameter estimates. Nevertheless, their approach is not applicable in our case since we are not in the framework of ergodic hidden Markov models and data sets from plant growth models are usually of small size.

Unfortunately, any stochastic EM-type algorithm that can be designed for the hidden Markov model given by Proposition 1 leads to a non-explicit M-step as well. For this reason, a numerical maximization procedure of quasi-Newton type, should be implemented at each iteration. For a large dimensional parameter vector this could be rather time consuming and may lead to convergence towards local maxima. [27] propose to separate a complicated M-step into smaller, more tractable conditional M-steps and update step by step the parameters of the model. This is the principle of the ECM (Expectation Conditional Maximization) algorithm, which is an interesting extension of the classical EM-algorithm. The ECM can also be used when we consider stochastic variants of the EM. In our setting we can benefit from this principle since we can reduce the number of parameters to be updated via numerical maximization by updating explicitly in a first CM (conditional maximization) step a large number of easily tractable parameters with fixed values of the rest. We describe this idea in the deterministic setting for notational simplicity since in the stochastic case we just have to replace all the smoothing functionals with their estimated counterparts.

Let $\mathcal{Q}(\theta;\theta') \stackrel{\text{def}}{=} \mathbb{E}_{\theta'}\big[\log p_\theta(Q_{0:N}, y_{0:N})|y_{0:N}\big]$ and assume that $\theta = (\theta_1, \theta_2)$ and marginal maximization w.r.t. θ_1 or θ_2 is easy. Then, the i-th iteration of the ECM algorithm consists in the following steps:

- E-step: Compute $\mathcal{Q}(\theta_1, \theta_2; \theta_1^{(i)}, \theta_2^{(i)})$
- CM-step:

$$\begin{aligned}\theta_1^{(i+1)} &= \arg\max_{\theta_1} \mathcal{Q}(\theta_1, \theta_2^{(i)}; \theta_1^{(i)}, \theta_2^{(i)}),\\ \theta_2^{(i+1)} &= \arg\max_{\theta_2} \mathcal{Q}(\theta_1^{(i+1)}, \theta_2; \theta_1^{(i)}, \theta_2^{(i)}).\end{aligned}$$

Recall that the parameter of the GreenLab model $\theta = (p, \sigma^2, \Sigma)$, where σ^2 and Σ are variance parameters related to model uncertainty and measurement errors respectively and $p = (\mu, s^{pr}, k_B, p_{al})$, where the first three are parameters related to the production equation and p_{al} to the allocation pattern. We will show that an ECM algorithm can be applied to this problem to reduce the complexity of the maximization problem, if we consider $\theta_1 = (\mu, \sigma^2, \Sigma)$, which is a 5-dimensional parameter vector (Σ depends on three independent parameters). The rest of the parameters form θ_2 and maximization is performed via the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton procedure.

Let us analyze the Q-function of the model. The density function of the complete model (complete likelihood function) by (8) and (9) is given by

$$p_\theta(q_{0:N}, y_{0:N}) = \prod_{n=1}^{N} p_\theta(q_n|q_{(n-T)^+:n-1}) \prod_{n=0}^{N} p_\theta(y_n|q_{n:(n+T-1)\wedge N}). \tag{15}$$

Let us denote $K_n \stackrel{\text{def}}{=} \mu^{-1} F_n$. In the rest, we identify the functions K_n and G_n (see (7)) with the induced random variable $K_n(\theta_2)$ and the induced random vector $G_n(\theta_2)$ respectively, for an arbitrary $\theta_2 \in \Theta_2$, where Θ_2 is an appropriate euclidean subset. By the definition of the Q-function, (15) and Proposition 1 we get

$$\mathcal{Q}(\theta;\theta') = \sum_{n=1}^{N} \mathbb{E}_{\theta'}\big[\log p_\theta(Q_n|Q_{(n-T)^+:n-1}) \mid y_{0:N}\big]$$

$$
\begin{aligned}
&+\sum_{n=0}^{N}\mathbb{E}_{\theta'}\left[\log p_\theta(y_n|Q_{n:(n+T-1)\wedge N}) \mid y_{0:N}\right]\\
&= C(\theta_2;\theta') + \mathcal{Q}_1(\mu,\sigma^2,\theta_2;\theta') + \mathcal{Q}_2(\Sigma,\theta_2;\theta'), \qquad (16)
\end{aligned}
$$

where

$$
\mathcal{Q}_1(\mu,\sigma^2,\theta_2;\theta') = -\frac{N}{2}\log\sigma^2 - N\log\mu - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\mathbb{E}_{\theta'}\left[\left(\mu^{-1}Q_nK_{n-1}^{-1}(\theta_2)-1\right)^2 \mid y_{0:N}\right], \qquad (17)
$$

$$
\begin{aligned}
\mathcal{Q}_2(\Sigma,\theta_2;\theta') &= -\frac{N+1}{2}\log(\det\Sigma)\\
&\quad -\frac{1}{2}\sum_{n=0}^{N}\mathbb{E}_{\theta'}\left[\big(y_n - G_n(\theta_2)\big)^\top\Sigma^{-1}\big(y_n - G_n(\theta_2)\big) \mid y_{0:N}\right], \qquad (18)
\end{aligned}
$$

and $C(\theta_2;\theta')$ is independent of θ_1.

Note that for fixed θ_2 the initial maximization problem of $\mathcal{Q}$ w.r.t. θ_1 can be separated into two distinct maximization problems of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ w.r.t. (μ,σ^2) and Σ respectively. In the following proposition we give the solution to the maximization problem. The proof is deferred to Appendix C.

Proposition 2. *Let* $\widehat{\theta}_{1,N}(\theta_2;\theta') = \left(\widehat{\mu}_N(\theta_2;\theta'), \widehat{\sigma^2_N}(\theta_2;\theta'), \widehat{\Sigma}_N(\theta_2;\theta')\right)$ *be the maximizers of the Q-function given by (16) when* θ_2 *is fixed. The update equations for* θ_1 *are given as follows:*

$$
\widehat{\mu}_N(\theta_2;\theta') = N^{-1}\sum_{n=1}^{N}\mathbb{E}_{\theta'}\left[Q_nK_{n-1}^{-1}(\theta_2) \mid y_{0:N}\right], \qquad (19)
$$

$$
\widehat{\sigma^2_N}(\theta_2;\theta') = N^{-1}\,\widehat{\mu}_N^{-2}(\theta_2;\theta')\sum_{n=1}^{N}\mathbb{E}_{\theta'}\left[Q_n^2K_{n-1}^{-2}(\theta_2) \mid y_{0:N}\right] - 1, \qquad (20)
$$

$$
\widehat{\Sigma}_N(\theta_2;\theta') = (N+1)^{-1}\sum_{n=0}^{N}\mathbb{E}_{\theta'}\left[\big(y_n - G_n(\theta_2)\big)\big(y_n - G_n(\theta_2)\big)^\top \mid y_{0:N}\right] \qquad (21)
$$

Remark 3. *In case* σ^2 *is fixed (assumed level of uncertainty), we get a different update equation for the parameter* μ*, given by*

$$
\widehat{\mu}_N(\theta_2;\theta') = \left(2N\sigma^2\right)^{-1}\left(\Delta^{1/2}(\theta_2;\theta') - \sum_{n=1}^{N}\mathbb{E}_{\theta'}\left[Q_nK_{n-1}^{-1}(\theta_2) \mid y_{0:N}\right]\right), \qquad (22)
$$

where

$$
\Delta(\theta_2;\theta') = \left(\sum_{n=1}^{N}\mathbb{E}_{\theta'}\left[Q_nK_{n-1}^{-1}(\theta_2) \mid y_{0:N}\right]\right)^2 + 4N\sigma^2\sum_{n=1}^{N}\mathbb{E}_{\theta'}\left[Q_n^2K_{n-1}^{-2}(\theta_2) \mid y_{0:N}\right].
$$

For a proof of (22) see Appendix C, Remark 4.

The detailed description of the E-step of the proposed algorithm and the recursive update of all the aforementioned smoothing functionals needed for the first explicit conditional maximization step is given in Appendix A.

5 Results and some practical considerations

In this section we present selected results which concern a simulated case (synthetic example) and a real data scenario from the sugar-beet. Several practical issues concerning the implementation of the proposed algorithm are also discussed. The synthetic example is chosen on purpose in the spirit of the application with the sugar-beet in order to show the theoretical performance of the algorithm when data are really generated from the assumed model. The state estimation part was implemented with the Improved filter, since by construction it is more informative than the blind filter which does not take into account the data in the proposal distributions.

5.1 Simulated data case

In this example we assume that the plant is cut after 50 growth cycles ($N = 50$). We illustrate the performance of the SISRECM algorithm that we developed by using a data file generated by a specific parameter file assumed to be the true one. The parameters are divided into two categories, those which are assumed to be known or calibrated directly in the field and the unknown parameter θ that has to be estimated. The values that we used to generate the data are given in Appendix D (Table 1). The update equations are given in Proposition 2 and equation (22) for a fixed σ_Q. The estimation of the covariance matrix of the measurement errors corresponds to the estimation of the standard deviations σ_b and σ_p (for measurements of the blade and the petiole respectively) and the correlation coefficient ρ. In order to estimate the parameters of the model we used the averaging technique (see, e.g., [6], page 407) which smooths the final estimates by averaging (weighted average) successively after a burn-in period all subsequent EM-estimates with weights proportional to the Monte-Carlo sample size used in the corresponding EM iterations (see Figure 1). In Table 1 we give the mean parameter estimates that we obtained from independent runs of the algorithm and under two different initial conditions (the standard deviations are also given).

Table 1: Parameter estimation results based on the synthetic example (see also Appendix D, Table 1). Estimates with the fully observed model are given in column 3. We give the means and the standard deviations of the estimates based on 50 independent runs and two different initializations (init1=real).

param.	real	fully-obs.	init2	mean1	mean2	std1	std2
a_b	3	2.909	2.4	2.793	2.795	0.0094	0.0103
a_p	3	2.905	2.4	2.818	2.820	0.0093	0.0102
P_p	0.8165	0.8152	0.6532	0.8150	0.8150	8×10^{-6}	9×10^{-6}
μ^{-1}	140	140.07	112	142.05	142.02	0.109	0.124
σ_b	0.1	0.1136	0.1	0.1187	0.1186	0.0001	0.0001
σ_p	0.1	0.1058	0.1	0.1102	0.1102	0.0001	0.0001
ρ	0.8	0.8158	0.8	0.8318	0.8318	0.0003	0.0003

Since our algorithm is stochastic we performed independent runs in order to increase the level of accuracy. The number of runs should be chosen by the user depending on the desired level. With 50 independent runs we can see very small differences in the estimated means (see Table 1, columns mean1 and mean2) since independent runs of the algorithm reduce the variance proportionally to the number of runs. For each independent run of the algorithm, the number of Monte-Carlo sample size was increased piecewise linearly for the first 50 iterations (starting from

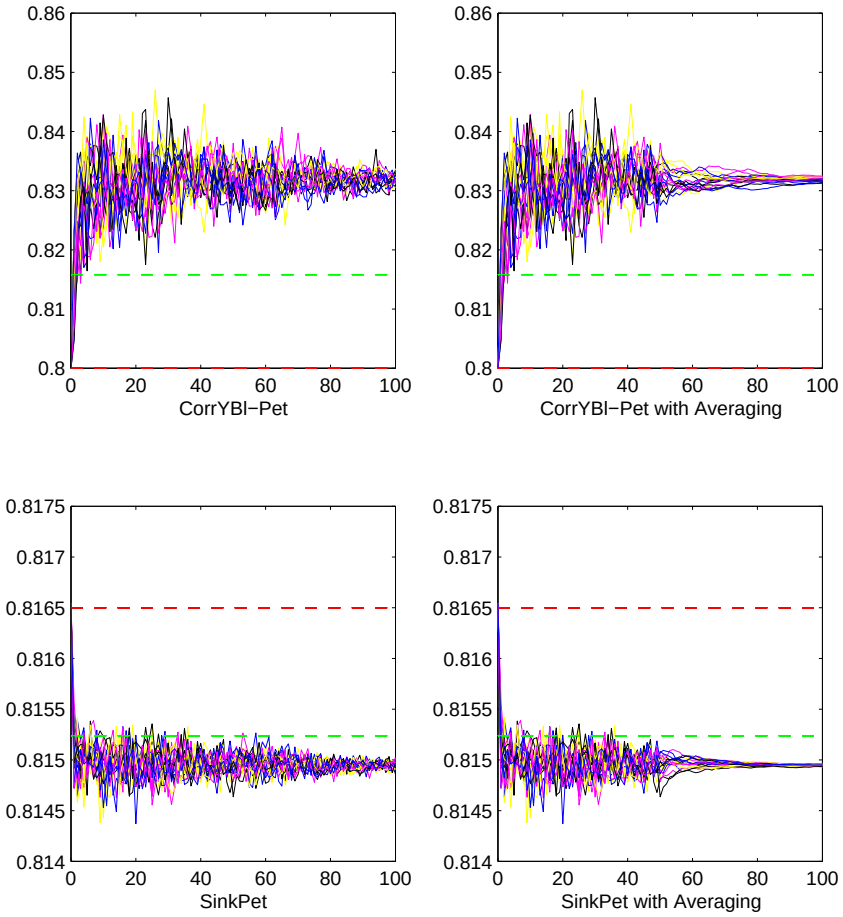

Figure 1: Parameter estimation during 100 EM iterations for the correlation coefficient and the sink petiole with 20 independent runs of the algorithm. The effect of averaging is shown in the right figures. The dotted lines which are nearest to the estimates correspond to the MLE if the data were fully observed and the other dotted lines to the parameters used to generate the data.

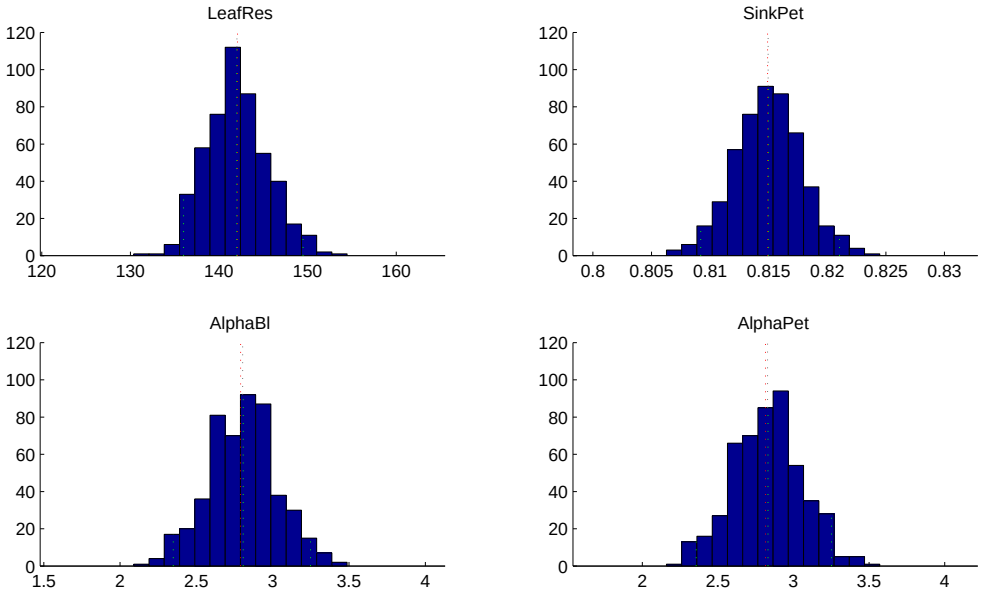

Figure 2: Bootstrap estimation of the marginal distributions of the MLE of the genetic parameters (Leaf resistance, Sink petiole, alpha blade and alpha petiole) based on a 500 bootstrap sample. The vertical dotted lines correspond to the estimated (0.025, 0.5, 0.975) quantiles of the MLE. In each subfigure, near to the median the supplementary vertical dotted line represents the estimated mean of the MLE

250, then increasing by 10 for the first 25 iterations and by 20 for the subsequent 25 iterations), and for the last 50 iterations we used a quadratic increase until we reach 10.000 trajectories. The algorithm stopped deterministically after 100 EM-iterations. For an automatic criterion the averaging technique can be used also as a way to impose a stopping criterion.

We used parametric bootstrap (see [35], Section 3.6.2.) for giving an approximate distribution of the MLE of the parameters. The bootstrap M-sample ($M = 500$ in our case) was formed by obtaining parameter estimates from data files generated independently from the estimated parameter mean1 given in Table 1. Each bootstrap estimate was based on a single run of the algorithm. For estimating with better precision higher moments or the tail of the distribution the bootstrap sample should be further increased and each individual bootstrap estimate should be taken as the mean of a certain number of independent runs of the algorithm. Nevertheless, depending on the desired precision this could prove to be rather computationally inefficient. In Figure 2 we give bootstrap estimates of the marginal distributions of the MLE of the genetic parameters. In Table 2 we give the estimates of the means and the standard deviations of both the MLE of the true and the complete data model (if the data were fully observed). Notice the difference in the standard deviations of the alpha parameters between the fully observed and the hidden model. These two parameters are mostly affected by the additional variability due to the incompleteness of the model (compare with the other parameters). The common behavior of the alpha parameters can be explained by the extremely high positive correlation of their estimators (see Table 4).

5.2 Experimental data from the sugar-beet

The experimental protocol carried out in 2006 in order to collect the experimental data from the sugar-beet is detailed in [23]. In this study, we tackle the problem of parameter estimation

Table 2: Parametric bootstrap estimates of the means and the standard deviations of the MLE of the hidden model and the MLE of the fully observed (FO) model for comparison reasons. The estimates are based on a 500 bootstrap sample and are MLE estimates obtained from 500 data files generated with true parameter the MLE given in column 2.

param.	MLE	FO-mean	MLE-mean	FO-std	MLE-std
a_b	2.793	2.792	2.804	0.078	0.224
a_p	2.818	2.816	2.828	0.084	0.226
P_p	0.815	0.815	0.815	0.0029	0.0029
μ^{-1}	142.05	142.03	142.03	2.245	3.425
σ_b	0.119	0.114	0.113	0.012	0.015
σ_p	0.110	0.106	0.105	0.012	0.013
ρ	0.832	0.833	0.828	0.047	0.053

based on a single plant. In this real-data scenario we make the assumption that blades and petioles have common expansion durations $T = 10$.

The parameters are divided into two categories, those which were calibrated directly in the field and the unknown parameter θ that has to be estimated.

First, based on this real-data scenario, we motivate our choice to introduce the HMM given by Proposition 1 to fit these data. As a by-product of its formulation this model introduces three types of correlations, i) in the masses of blade and petiole preformed at the same GC, ii) in the masses of blades preformed at different GCs (correlation in time), iii) in the masses of petioles preformed at different GCs (correlation in time). The first correlation is directly taken into account by including the correlation coefficient as a parameter to be estimated. Correlation in time is indirectly introduced in the other two cases by introducing stochasticity (with $T > 1$) in the biomass production. Intuitively, the simpler model described in Section 2-iii) would be preferable to the new one if these correlations were not statistically significant. For the rest, the simpler model is referred to as the normal model. Let us now denote by r_{bp} the sample correlation corresponding to case i), and r_b and r_p, the first-order sample autocorrelations corresponding to case ii) and iii) respectively, if mean theoretical masses are estimated by their fitted means based on the normal model estimates. The results that we obtained are given as follows:

$$(r_{bp},\ r_b,\ r_p) = (0.0820,\ 0.6318,\ 0.6606). \tag{23}$$

Note that in the classical test of the correlation coefficient, under the null hypothesis of null correlation, with significance level 0.05 the critical values are 0.304 and 0.308 for the first and the last two correlations respectively. This shows that the fitting with the normal model reveals that no significant correlation is present in the mass measurements of the blade and petiole but the correlation in time for blade and petiole are significant and cannot be neglected. This is reflected in the model comparison that we make in the sequel.

Now, we illustrate the performance of the SISRECM algorithm based on this experimental data file. The parameter σ^2 represents a standard level of uncertainty ($\sigma = 0.1$) for the mean biophysical model given by (5). We used several initializations and the results are similar. In Appendix D (Table 2) we give the initial values for the presented results. The simulation procedure that we used is the same as explained in the synthetic example. In Table 3 we give the MLE that we obtained as a mean of 50 independent runs and the estimates of the means and the standard deviations of the distribution of the MLE. We also indicate the root mean square errors with respect to the MLE and the coefficients of variation. Note that the parameter with

Table 3: Parameter estimates based on the SISRECM algorithm and bootstrap estimates of the means and the standard deviations of the marginal distributions of the MLE based on a 500 bootstrap sample. The root mean square errors (RMSE) and the coefficients of variation (CV) are also given.

param.	MLE	MLE-mean	MLE-std	RMSE	CV
a_b	2.829	2.794	0.280	0.282	0.100
a_p	1.823	1.805	0.225	0.225	0.125
P_p	0.815	0.814	0.016	0.016	0.020
μ^{-1}	98.125	99.05	5.357	5.432	0.054
σ_b	0.076	0.073	0.008	0.009	0.110
σ_p	0.059	0.056	0.007	0.007	0.125
ρ	0.1266	0.137	0.161	0.162	1.175

the highest coefficient of variation (compare std/mean) is as expected the correlation coefficient of the measurement error model (since the range of its values is close to zero and takes also negative values). All the MLE of the parameters in the real data scenario have larger estimated standard deviations than the ones in the synthetic example. The bootstrap estimates of the marginal distributions of the MLE corresponding to the genetic parameters as well as estimated confidence intervals are given in Appendix D (Figure 1). In Table 4 we give the estimated correlation matrix of the MLE (the lower diagonal matrix). Notice the very high positive

Table 4: Bootstrap estimation of the correlation matrices of the MLE based on a 500 bootstrap sample. The upper diagonal matrix gives the estimated correlation coefficients for the synthetic example and the lower diagonal matrix for the real data case.

param.	a_b	a_p	P_p	μ^{-1}	σ_b	σ_p	ρ
a_b	1	0.970	-0.005	-0.783	-0.074	0.003	-0.036
a_p	0.941	1	-0.031	-0.741	-0.057	0.013	-0.021
P_p	-0.055	-0.013	1	-0.016	0.070	0.025	-0.007
μ^{-1}	-0.936	-0.822	0.060	1	0.082	0.020	0.028
σ_b	-0.008	-0.013	0.070	-0.004	1	0.6948	0.600
σ_p	-0.021	-0.005	-0.003	0.028	0.038	1	0.630
ρ	0.101	0.092	-0.018	-0.122	0.081	0.104	1

correlation between the MLE of the alpha parameters and their joint highly negative correlation with the MLE of the leaf resistance (the inverse of the leaf transpiration efficiency).

Finally, driven by the results and the discussion following (23) we repeated the fitting with the HMM under the assumption that $\rho = 0$ (one parameter less to estimate). The three competing models: i) the normal model, ii) the HMM1 (ρ as a free parameter) iii) the HMM2 ($\rho = 0$) were compared for their fitting quality on the basis of the two most commonly used model selection citeria, the corrected Akaike information criterion (corrected AIC) and the Bayesian information criterion (BIC). In Table 5 we present the results of this comparison. The best model according to both criteria was shown to be the HMM with $\rho = 0$ (attains the minimum AICc and BIC). Even if HMM1 has the maximum log-likelihood (see Table 6) the penalty term for the estimation of one additional parameter makes it the least probable. These results are consistent with the preliminary tests based on the sample correlation coefficients given in (23).

The MLE corresponding to all competing models are given in Table 6, followed by bootstrap estimated 95% confidence intervals of the best model.

Table 5: Corrected AIC and BIC evaluation for the three competing models: i) the normal model, ii) the HMM1 with the correlation coefficient ρ as a free parameter and iii) the HMM2 with no correlation ($\rho = 0$). The values in parenthesis correspond to the estimated standard deviation of the mean values of AICc and BIC for HMM1 and HMM2 based on 100 samples of 5×10^5 independent evaluations. For completeness: $AICc = -2(\log \hat{L} - d) + 2d(d+1)/(n-d+1)$ and $BIC = -2 \log \hat{L} + d \log n$, where d is the number of free parameters and n the sample size

model	normal	HMM1	HMM2
AICc	-343.319	-342.174 (0.021)	-344.022 (0.035)
BIC	-329.825	-326.631 (0.021)	-330.528 (0.035)

Table 6: MLE obtained with the normal, the HMM1 (ρ as a free parameter) and the HMM2 (ρ=0) model for the sugar-beet data set, together with their estimated log-likelihoods (up to a constant term). For the sample size see Table 5. The last two lines give 95% confidence intervals for the best model HMM2, based on a 500 bootstrap sample.

model	a_b	a_p	P_p	μ^{-1}	σ_b	σ_p	ρ	$\log \hat{L}$
normal	2.956	2.030	0.8162	96.33	0.081	0.061	0	178.205 (0.000)
HMM1	2.829	1.822	0.8147	98.13	0.076	0.059	0.13	178.824 (0.010)
HMM2	2.868	1.858	0.8145	98.03	0.074	0.059	0	178.556 (0.018)
$q_{0.025}$	2.343	1.421	0.7802	88.89	0.056	0.043	0	-
$q_{0.975}$	3.398	2.329	0.8462	109.58	0.088	0.071	0	-

6 Discussion

In this paper, we developed a theoretical framework for describing plant growth in a class of plants characterized by deterministic structural development. We also proposed a stochastic variant of the ECM algorithm for parameter estimation of this class of models. The results obtained with the sugar-beet plant are encouraging but further research effort is needed in order to improve modeling and parameter estimation both at individual and population levels. Further improvements should be made to the existing approach so as to take into account in a more informative way the organ masses of the last immature members for large expansion durations. Finding parameterizations which are at the same time adequate and parsimonious in the modeling of living organisms is always a difficult and demanding task. The difficulty increases given the limited life time or observation time of several plants.

One implementation issue that still attracts research interest concerns the Monte-Carlo sample augmentation strategy. The simulation schedule that we used is deterministic (see Section 5) and is similar to that of [6] (see Section 11.1.2). From our experience, even with small sample sizes at the first iterations, parameter estimates are driven rapidly towards the region of interest. However, a lot of simulation effort is needed near convergence to avoid "zig-zagging". A small quadratic increase of the sample size w.r.t. the number of ECM steps was rather satisfying to keep a moderate final sample size in a sufficient number of ECM iterations allowing to

detect convergence. The averaging technique further improved the precision of the estimates. Despite the simplicity and the good control that one can have with a deterministic schedule, in the routine use of the method, automated and data-driven algorithms can offer an interesting alternative (see, e.g., [5]). Further studies are needed to assess the benefits of such automated procedures.

To our knowledge, convergence results are available only for the deterministic ECM ([27]) and not for its stochastic counterpart. The difficulty of handling these questions increases since one conditional maximization step is performed numerically (see also [22]). The strongest known results for MCEM algorithms are those of [14] concerning only (curved) exponential parametric families. Convergence analysis of the present algorithm is a specific focus of our current research work.

Further perspectives include: i) the model validation in a large variety of plants with deterministic organogenesis. For this purpose, different types of model and measurement errors should be tested and compared with the help of model selection techniques, ii) alternative simulation based parameter estimation techniques could be developed, tested and compared with the existing method, as for example an ECM algorithm based on a Markov Chain Monte-Carlo state estimation technique, iii) a generalization of the proposed methodology for plants with stochastic organogenesis (including trees), where the total number of organs of each class at each Growth Cycle are random variables, iv) a natural extension of the proposed methodology at population level, which seems to be feasible and realistic for low inter-individual variability.

Supplementary materials

Appendices are available online, giving a more detailed description of the algorithm, proofs of results and figures & tables referenced in the text.

Acknowledgements

The authors are grateful to the editor-in-chief, the associate editor and the referees. Their valuable comments and suggestions improved considerably this paper.

References

[1] D. Barthélémy and Y. Caraglio. Plant architecture: a dynamic, multilevel and comprehensive approach to plant form, structure and ontogeny. *Annals of Botany*, 99(3):375–407, 2007.

[2] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[3] A. Björk. *Numerical methods for least squares problems*. SIAM, 1996.

[4] R. Boyles. On the convergence of EM algorithm. *Journal of the Royal Statistical Society*, 45:47–50, 1983.

[5] B. S. Caffo, W. Jank, and G. L. Jones. Ascent–based Monte Carlo Expectation–Maximization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:235–251, 2005.

[6] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, New York, 2005.

[7] G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82, 1985.

[8] P.-H. Cournède. Dynamic system of plant growth. HDR Thesis, Univ. of Montpellier II, 2009.

[9] P.-H. Cournède, M.Z. Kang, A. Mathieu, J.-F. Barczi, H.P. Yan, B.G. Hu, and P. de Reffye. Structural factorization of plants to compute their functional and architectural growth. *Simulation*, 82(7):427–438, 2006.

[10] P.-H. Cournède, V. Letort, A. Mathieu, M.-Z. Kang, S. Lemaire, S. Trevezas, F. Houllier, and P. de Reffye. Some parameter estimation issues in functional-structural plant modelling. *Mathematical Modelling of Natural Phenomena*, 6(2):133–159, 2011.

[11] P. de Reffye and B.G. Hu. Relevant choices in botany and mathematics for building efficient dynamic plant growth models: the greenlab case. In B.G. Hu and M. Jaeger, editors, *Plant Growth Models and Applications*, pages 87–107. Tsinghua Univ. Press and Springer, 2003.

[12] B. Delyon, V. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27:94–128, 1999.

[13] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 39:1–38, 1977.

[14] G. Fort and E. Moulines. Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics*, (31):1220–1259, 2003.

[15] N. Gordon, D. Salmond, and A. F. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings-F, Radar Signal Process.*, 140(2):107–113, 1993.

[16] Y. Guo, Y.T. Ma, Z.G. Zhan, B.G. Li, M. Dingkuhn, D. Luquet, and P. de Reffye. Parameter optimization and field validation of the functional-structural model greenlab for maize. *Annals of Botany*, 97:217–230, 2006.

[17] J. Handschin and D. Mayne. Monte carlo techniques to estimate the conditional expectation in multistage nonlinear filtering. *Intern. Journal of Control*, 9(5):547–559, 1969.

[18] J. Hull and A. White. The pricing of options on assets with stochastic volatilities. *J. Finance*, 42:281–300, 1987.

[19] W. Jank. Stochastic Variants of EM: Monte Carlo, Quasi–Monte Carlo and More. In *Proceedings of the American Statistical Association*, 2005.

[20] A. Jullien, A. Mathieu, J.-M. Allirand, A. Pinet, P. de Reffye, P.-H. Cournède, and B. Ney. Characterisation of the interactions between architecture and source:sink relationships in winter oilseed rape (brassica napus l.) using the greenlab model. *Annals of Botany*, 107(5):765–779, 2011.

[21] G. Kitagawa and S. Sato. Monte carlo smoothing and self-organising state-space model. In A. Doucet, de Freitas N., and Gordon N., editors, *Sequential Monte Carlo Methods in Practice*, pages 178–195. Springer, New York, 2001.

[22] K. Lange. A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):425–437, 1995.

[23] S. Lemaire, F. Maupas, P.-H. Cournède, and P. de Reffye. A morphogenetic crop model for sugar-beet (*beta vulgaris l.*). In *International Symposium on Crop Modeling and Decision Support: ISCMDS 2008, April 19-22, 2008, Nanjing, China*, 2008.

[24] C. Loi and P.-H. Cournède. Generating functions of stochastic L-systems and application to models of plant development. *Discrete Mathematics and Theoretical Computer Science Proceedings*, AI:325–338, 2008.

[25] A. Mathieu, P.-H. Cournède, V. Letort, D. Barthélémy, and P. de Reffye. A dynamic model of plant growth with interactions between development and functional mechanisms to study plant structural plasticity related to trophic competition. *Annals of Botany*, 103(8):1173–1186, 2009.

[26] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons Inc., 2008.

[27] X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika,*, 80:267–278, 1993.

[28] J. Olsson, O. Cappé, R. Douc, and E. Moulines. Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179, 2006.

[29] R. Sievänen, E. Nikinmaa, P. Nygren, H. Ozier-Lafontaine, J. Perttunen, and H. Hakula. Components of a functional-structural tree model. *Annals of Forest Sciences*, 57:399–412, 2000.

[30] W. Taylor. Small sample properties of a class of two-stage Aitken estimator. *Econometrica*, 45(2):497–508, 1977.

[31] J. Warren-Wilson. Ecological data on dry matter production by plants and plant communities. In E.F. Bradley and O.T. Denmead, editors, *The collection and processing of field data*, pages 77–123. Interscience Publishers, New York, 1967.

[32] G. Wei and M. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.

[33] C. Wu. On the convergence properties of EM algorithm. *The Annals of Statistics*, 11:95–103, 1983.

[34] V. Zaritskii, V. Svetnik, and L. Shimelevich. Monte-Carlo techniques in problems of optimal data processing. *Autom. Remote Control*, 12:2015–2022, 1975.

[35] W. Zucchini and I.L. MacDonald. *Hidden Markov Models for Time Series - An Introduction Using R*. Chapman and Hall, 2009.